DOCUMENT RESUME

ED 152 830 95 TM 007 031

AUTHOR Novick, Melvin R.; Lewis, Charles
TITLE Prescribing Test Length for Criterion-Referenced Measurement. I. Posttests. ACT Technical Bulletin No. 18.
INSTITUTION American Coll. Testing Program, Iowa City, Iowa. Research and Development Div.
SPONS AGENCY Office of Education (DHEW), Washington, D.C.
PUB DATE Jan 74
GRANT OEG-0-72-0711
NOTE 35p.

EDRS PRICE MF-$0.83 HC-$2.06 Plus Postage.
DESCRIPTORS Behavioral Objectives; *Criterion Referenced Tests; Cutting Scores; Elementary Secondary Education; *Expectancy Tables; *Guidelines; Individualized Instruction; *Mastery Learning; *Mastery Tests; Performance; Post Testing; Sampling; Student Evaluation; *Test Construction
IDENTIFIERS *Test Length

ABSTRACT

In a program of Individually Prescribed Instruction (IPI), where a student's progress through each level of a program of study is governed by his performance on a test dealing with individual behavioral objectives, there is considerable value in keeping the number of items on each test at a minimum. The specified test length for each objective must, however, be adequate to provide sufficient information regarding the student's degree of mastery of the behavioral objective being tested. Just what the minimum acceptable length will be depends on the manner in which test information is used to make decisions about individual students, the level of functioning required for defining mastery of an objective, the relative losses incurred in making false positive and false negative decisions, the background information available on the student and on the instructional process, and the premium on testing time within the instructional process. Some broad guidelines regarding test length of IPI posttests are included. A number of tables present data on the probability of the students achieving mastery level. (Author/MV)

TM

ACT TECHNICAL BULLETIN NO. 18

PRESCRIBING TEST LENGTH FOR CRITERION-REFERENCED MEASUREMENT

by

Melvin R. Novick and Charles Lewis

The American College Testing Program
and
The University of Iowa

The University of Illinois

The Research and Development Division

The American College Testing Program

P. O. Box 168, Iowa City, Iowa 52240

January, 1974

Prescribing Test Length For Criterion-Referenced Measurement*

I. Posttests

by

Melvin R. Novick and Charles Lewis

The American College Testing Program and The University of Iowa

The University of Illinois

Introduction

In a program of Individually Prescribed Instruction (IPI), where a student's progress through each level of a program of study is governed by his performance on a test dealing with individual behavioral objectives, there is considerable value in keeping the number of items on each test at a minimum. The specified test length for each objective must, however, be adequate to provide sufficient information regarding the student's degree of mastery of the behavioral objective being tested. Just what the minimum acceptable length will be depends on the manner in which test information is used to make decisions about individual students, the level of functioning required for defining mastery of an objective, the relative losses incurred in making false positive and false negative decisions, the background information available on the student and on the instructional process, and the premium on testing time within the instructional process. Our purpose in

---

*The research reported herein was performed pursuant to Grant No. OEG-0-72-0711 with the Office of Education, U. S. Department of Health, Education, and Welfare, Melvin R. Novick, Principal Investigator. Contractors undertaking such projects under Government sponsorship are encouraged to express freely their professional judgment in the conduct of the project. Points of view or opinions stated do not, therefore, necessarily represent official Office of Education position or policy. We are grateful to Charles Davis and Nancy Petersen for helpful comments and computations. This paper will be published in the GSE Monograph Series in Evaluation, Number 3, a publication supported in part by the National Institute of Education and by the American Educational Research Association.

this paper is to discuss these issues and provide some broad guidelines for test-length specification for IPI posttests. These specifications will be tentative because of unresolved substantive and methodological issues, but we believe that they should provide some improvement on current practice. A separate, and rather more complex treatment will be required for placement and pretest length specification.

Background

In a criterion-referenced measurement approach to Individually Prescribed Instruction, we imagine a population of test items, having mixed item difficulty, dealing with a particular objective and an ideal decision which advances a student past this objective if he is able to answer at least a given percentage of the items in the population. This minimum passing percentage, the so-called criterion level, simply reflects the degree of mastery deemed sufficient for this objective (although it implicitly involves the difficulty of the items as well). The actual percentage of items that a person would answer correctly in the population of items is called his level of functioning. In practice, the advancement-retention decision must be made from a small sample of observations (test items), and, hence, errors in the decision process must be expected.

One common treatment of the test length problem in a criterion-referenced measurement context has been given by Millman (1972). He studied a standard decision rule which advances the student if the percent of items correctly answered on a test equals or exceeds the required criterion level. Here it is assumed that the items on the test may be treated as a random sample from the population of interest, so that the obtained percentage correct is a useful estimate of the true population percentage for the student. Using binomial probability

Table 1

Percent of Students Expected To Be Incorrectly Advanced or Retained

Specified Criterion Level .70

Advancement Score	No. of Test Items	Student's True Level of Functioning* 50	55	60	65	70	75	80	85	90	95
6	7	6	10	16	23	67	55	42	28	15	4
6	8	15	22	32	43	45	32	20	11	4	1
7	9	9	15	23	34	54	40	26	14	5	1
7	10	17	27	38	51	35	22	12	5	1	-
8	11	11	19	30	43	43	29	16	7	2	-
9	12	7	13	23	35	51	35	20	9	3	-
10	13	5	9	17	28	58	42	25	12	3	-
11	14	3	6	12	22	64	48	30	15	4	-
12	15	2	4	9	17	70	54	35	18	6	-

Specified Criterion Level .75

Advancement Score	No. of Test Items	Student's True Level of Functioning* 50	55	60	65	70	75	80	85	90	95
6	8	15	22	32	43	55	32	20	11	4	1
7	9	9	15	23	34	46	40	26	14	5	1
8	10	6	10	17	26	38	47	32	18	7	1
9	11	3	7	12	20	31	55	38	22	9	2
9	12	7	13	23	35	49	35	20	9	3	-
16	20	1	2	5	12	24	58	37	17	4	-
17	21	-	1	4	9	20	63	41	20	5	-
18	22	-	1	3	7	17	68	46	23	6	-

Table 1 (continued)

Specified Criterion Level .80

Advancement Score	No. of Test Items	Student's True Level of Functioning* 50	55	60	65	70	75	80	85	90	95
6	7	6	10	16	23	33	45	42	28	15	4
7	8	4	7	11	17	26	37	50	34	19	6
8	9	2	4	7	12	20	30	56	40	23	7
8	10	6	10	17	26	38	53	32	18	7	1
9	11	3	7	12	20	31	46	38	22	9	2
10	12	2	4	8	15	25	39	44	26	11	2
11	13	1	3	6	11	20	33	50	31	13	2
12	15	2	4	9	17	30	46	35	18	6	-
17	20	-	1	2	4	11	23	59	35	13	2
19	22	-	-	1	3	7	16	67	42	17	2

Specified Criterion Level .85

Advancement Score	No. of Test Items	Student's True Level of Functioning* 50	55	60	65	70	75	80	85	90	95
7	8	4	7	11	17	26	37	50	34	19	6
8	9	2	4	7	12	20	30	44	40	23	7
9	10	1	2	5	9	15	24	38	46	26	9
10	11	1	1	3	6	11	20	32	51	30	10
11	12	-	1	2	4	9	16	28	56	34	12
17	19	-	-	1	2	5	11	24	56	29	7
19	21	-	-	-	1	3	8	18	63	35	8

*The true level of functioning is the percent of items a student would be able to answer correctly if he were given the entire universe of items.

Students having true level of functioning values less than the specified criterion level should fail a test composed of all items from this universe. However, on any given test of finite length, some of these students will get more than the minimum advancement percent of the items correct and be considered as "passers". The expected percent of such incorrect advancements are given in the body of the table to the left of the dotted line.

Students having true level of functioning values equal to or greater than the minimum advancement percent should pass such a test. The percent of these students who will be incorrectly retained are shown in the table to the right of the dotted line.

tables, Millman obtained the probability that a student with a given true level of functioning would be incorrectly advanced or retained by this procedure.

Table 1 expands on some of Millman's computations and gives the conditional probability of incorrect advancement or retention for a variety of true levels, test lengths, and minimum passing percentages. The first impression this table provides is that a substantial proportion (sometimes more than half) of the students with true levels close to, or at the criterion level, will be incorrectly advanced or retained, at least for the test lengths considered. There appears to be a slight improvement in accuracy of decision as the test length increases from 8 to 22 items, although this effect is largely hidden by fluctuation in the probabilities, due to changes in the percentage correct required for advancement. For example, with a criterion level of .7, the percentage correct required for advancement is .75, .78, .70, .73, or .75 for test lengths of 8, 9, 10, 11, or 12 items, respectively. This brings up a question as to the optimality of the decision procedure assumed in Table 1. To provide a framework for answering this question, let us consider some of the issues involved.

Suppose seven out of eight were taken as the minimum advancement score when the criterion level is .75; the probability of incorrect advancement would decrease substantially for all students with true levels below the criterion level. This is shown in Table 2. On the other hand, those above .75 suffer a substantial increase in their chances of being incorrectly retained. Apparently, a more general framework is required before even the decision procedure can be chosen, much less any judgment made concerning minimum test length. This framework would need to take into account on which side of .75 small expected errors were considered to be more important.

Table 2

Percent of Students Expected To Be Incorrectly Advanced or Retained

Criterion Level = .75 Test Length = 8

Advancement Score	True Level 50	55	60	65	70	75	80	85	90	95
6	15	22	32	43	55	32	20	11	4	1
7	4	7	11	11	26	63	50	34	19	6

A Framework For Specifying Test Length

Table 1 is very helpful in identifying the seriousness of the problem of short tests. From a practical point of view, however, a solution to the problem must involve looking at a different conditional probability, and abandoning the simple decision procedure that Millman has so convincingly demonstrated to be inadequate. Instead of the probability that a student will attain a particular test score, given his true level, it is the probability that a student's true level of functioning exceeds the specified criterion level, given his test score, which is required in making a decision. In other words, it is the test score--not the true level--which is given (i.e. observed), and which is the basis for any decision to advance or retain the student. Thus, a student should be advanced only if the probability that he has attained or surpassed the criterion level, given his test score, is sufficiently high. To obtain the necessary probability, an application of Bayes theorem is required. In such an analysis, prior knowledge (expressed in probabilistic terms) of the student's true level of functioning is combined with the (binomial) model information relating the observed test score to true level; and, the result is a posterior probability

distribution for true level of functioning, given test score. The probability this distribution assigns to levels above the criterion is the quantity of interest. In this formulation, the problem can be described as selecting a minimum sample size and an advancement score, so that students attaining that score will then have a sufficiently high probability of having at least the minimum required level of functioning.

As a first approximation, let us suppose our knowledge of a student's true level of functioning is vague, prior to having his test results. If this state of knowledge is characterized by selecting a uniform distribution on the interval from zero to unity for true level, π, Bayes theorem provides the posterior probabilities listed in Table 3 for various scores and test lengths. The posterior distributions on which these probabilities are based all belong to the Beta family, and the parameters in each case are those given in the table, primarily for future reference.

To generate a decision procedure on the basis of Table 3, we must select a criterion level (π_o) and a minimum acceptable probability that a student's true level (π) exceeds this criterion. Thus, for example, we might take $\pi_o = .80$ and the minimum acceptable $\text{Prob}(\pi \geq \pi_o | x, n) = .50$, where x is test score and n is test length. We would then be saying that we wanted to advance the student only if we were at least 50% sure that his level of functioning was above .80. Then, using Table 3, we see that with n = 8, all students having $x \geq 7$ would advance to the next objective, but not those with $x = 6$. For a test of 12 items, the minimum advancement score would be 10 correct.

Note, however, that if we required 80% assurance that the true level of functioning was above .80, $[\text{Prob}(\pi \geq .80) \geq .80]$, then even those with eleven correct responses to twelve items would not be advanced. We think

Table 3

Probability Student's True Level Of Functioning Is Greater Than π_o Given A Uniform Prior Distribution

Minimum Advancement Score	No. of Test Items	Posterior Distribution	Criterion Level--π_o									
			50	55	60	65	70	75	80	85	90	95
6	8	β(7, 3)	91	85	77	66	54	40	26	14	5	1
7	8	β(8, 2)	98	96	93	88	80	70	56	40	23	7
8	8	β(9, 1)	100	100	99	98	96	92	87	77	61	37
7	9	β(8, 3)	95	90	83	74	62	47	32	18	7	1
8	9	β(9, 2)	99	98	95	91	85	76	62	46	26	9
9	9	β(10, 1)	100	100	99	99	97	94	89	80	65	40
7	10	β(8, 4)	89	81	70	57	43	29	16	7	2	-
8	10	β(9, 3)	97	93	88	80	69	54	38	22	9	2
9	10	β(10, 2)	99	99	97	94	89	80	68	51	30	10
8	11	β(9, 4)	93	87	77	65	51	35	21	9	3	-
9	11	β(10, 3)	98	96	92	85	75	61	44	26	11	2
10	11	β(11, 2)	100	99	98	96	92	84	73	56	34	12
9	12	β(10, 4)	95	91	83	72	58	42	25	12	3	-
10	12	β(11, 3)	99	97	94	89	80	67	50	31	13	2
11	12	β(12, 2)	100	100	99	97	94	87	77	60	38	14

that it is unreasonable to require perfect performance as a standard for advancement, and therefore, we need to improve upon this analysis. One way is to use a longer test, but we can, at least, hope to find a procedure in which a 12-item test will be adequate.

The results in Table 3, although they provide relevant information for mastery decisions about students based on test scores, do not take full advantage of the power which is available through the use of prior knowledge. In particular, it will seldom be the case that our knowledge of a student's true level is adequately described by a uniform distribution. For example, our prior probability that a student is functioning above a criterion level of .8 might be approximately .75. This would be the case if historical data suggested that about 75% of the students who completed a unit of Individually Prescribed Instruction proved to be at or above mastery level. Moreover, we might judge the strength of our knowledge to be roughly equivalent to that based on a score from a 12-item test. (A method for making this assesment will be referenced shortly.)

When working with a binomial model, it is convenient and generally very satisfactory to select a member of the Beta class of distributions to characterize prior beliefs (Novick and Jackson, 1974). If this is done, the posterior distribution is easily obtained, and in every instance will again be a member of the Beta family. In fact, if the prior distribution is $\beta(a, b)$ and x success in n trials are observed, then the posterior distribution is $\beta(x + a, n - x + b)$. This can be seen in Table 3, where it is noted that the uniform distribution is $\beta(1, 1)$. If we restrict ourselves to prior distributions in the Beta family, the beliefs specified in the previous paragraph are characterized by $\beta(10.254, 1.746)$. Given this prior

distribution and the indicated test results, the posterior distributions and posterior probabilities of exceeding various criteria are provided in Table 4. The precise stipulation of prior distributions must always be done carefully, but extensive aids (Novick and Jackson, 1974, Novick, Lewis, and Jackson, 1973) are available, and indeed an elaborate system of Computer Assisted Data Analysis (CADA) is available (Novick, 1973) to help an instructional decision maker specify his prior distribution. A yet more sophisticated way of getting prior and posterior distributions for each person is derived by Lewis, Wang, and Novick (1973) and the required tables are given by Wang (1973). For the present, we shall suppose that this work has been done carefully and that the prior distribution used in the construction of Table 4 is appropriate.

Tables 3 and 4 demonstrate clearly the impact of prior knowledge on our interpretation of test results. In Table 3, for example, the posterior probability that a student with a score of six out of eight items correct has a true level greater than .80 is only .26, whereas in Table 4 this probability has increased to .60. This result should not be surprising, in view of the fact that we have now set this probability to be .75, apriori as compared to .20 in Table 3. If we felt the chances to be very good that the student had mastered an objective (to a level above .8) before we saw the test results, then a score of six out of eight will not substantially change our beliefs; it will lower the probability, but aposteriori may still leave the odds in favor of mastery. In many applications, a prior probability of mastery may be no more than .60, but the results will still differ sharply from those obtained, assuming vague prior information. Note that if we were to adopt the rule that we will advance a student if the aposteriori probability of mastery is at least

Table 4

Probability Student's True Level of Functioning Is Greater Than π_o Given A $\beta(10.254, 1.746)$ Prior Distribution

Minimum Advancement Score	No. of Test Items	Posterior Distribution	Criterion Level--π_o									
			50	55	60	65	70	75	80	85	90	95
6	8	$\beta(16.254, 3.746)$	100	100	98	96	90	78	60	37	15	2
7	8	$\beta(17.254, 2.746)$	100	100	100	99	97	92	81	62	36	10
8	8	$\beta(18.254, 1.746)$	100	100	100	100	99	98	94	85	66	32
7	9	$\beta(17.254, 3.746)$	100	100	99	97	92	82	65	41	17	2
8	9	$\beta(18.254, 2.746)$	100	100	100	99	98	93	84	66	39	11
9	9	$\beta(19.254, 1.746)$	100	100	100	100	100	98	95	87	69	34
7	10	$\beta(17.254, 4.746)$	100	99	97	93	84	68	47	24	7	1
8	10	$\beta(18.254, 3.746)$	100	100	99	98	93	84	68	45	19	3
9	10	$\beta(19.254, 2.746)$	100	100	100	99	98	95	86	69	42	12
8	11	$\beta(18.254, 4.746)$	100	99	98	94	87	72	51	27	8	1
9	11	$\beta(19.254, 3.746)$	100	100	100	98	95	87	72	48	22	3
10	11	$\beta(20.254, 2.746)$	100	100	100	100	99	96	88	72	45	13
9	12	$\beta(19.254, 4.746)$	100	100	99	96	89	76	55	30	10	1
10	12	$\beta(20.254, 3.746)$	100	100	100	99	96	89	75	52	24	4
11	12	$\beta(21.254, 2.746)$	100	100	100	100	99	96	90	75	48	14

Note: The mean and mode, respectively of $\beta(10.254, 1.746)$ are .855 and .925 and for this distribution $\text{Prob}(\pi > \pi_o)$ for π_o = .70, .75, .80, .85 are .92, .86, .75, and .59, respectively. A close look at these distributional characteristics will help a decision maker determine if this prior distribution is a realistic characterization of his beliefs.

.50, then in this example, we will advance him if the prior distribution were that of Table 4, but not if it were that of Table 3.

When the decision maker specifies an informative prior distribution, he is saying, in effect, that he wants a decision which will have a high probability of being correct in that portion of the decision space in which he thinks the student's ability truly lies. For example, referring to Taole 2, a decision maker with a high prior probability that the student had a true level of functioning below .75 would, by virtue of his analysis, require a minimum passing score of seven correct out of eight items. This would assure him a low probability of misclassification for all values below .75. Another decision maker with high prior probability that the student was above criterion level would likely require only six out of eight correct, and thus have low probability of an incorrect decision for values of .75 or above.

Once we have decided to work with the posterior probability that a student's level of functioning exceeds some criterion, given his test score, and have made use of our prior knowledge in obtaining this probability, another issue remains to be settled before we can turn to the question of test length. Simply stated, we need to know how sure we should be that a student has mastered an objective at the chosen level before we make the decision to allow him to advance to the next objective. For instance, is a posterior probability of at least .5, as was used in the last example, a reasonable choice in all cases? Almost certainly this last question should be answered in the negative. The point at issue here comes down to an understanding of the relative disutilities or losses associated with the false positive and false negative errors.

If it were no more serious to advance a student whose level was below the criterion than to retain a student who was above, we would be behaving optimally if we were to advance students with posterior probabilities above .5 and retain the others. In many situations the prior probability will be this high, and hence an advancement decision could then be made on an apriori basis. On the other hand, we might consider the loss to be twice as great for a false advancement than for a false retention. In this case, we should only advance those students whose posterior probability for being above the criterion exceeds 2/3. The general result is that we shall achieve the smallest expected loss if we match the posterior odds to the loss ratio. Thus, if the loss ratio is 2 to 1 (false advance to false retain), a probability of 2/(2 + 1) gives matching odds of 2/3 to 1/3 above criterion to below criterion).

Table 5

Losses Associated With Incorrect Decisions

		True Level	
		$\pi \geq \pi_o$	$\pi < \pi_o$
Decision	Advance	0	a
	Retain	b	0

To express the result symbolically, consider the notation of Table 5. Here a is the loss associated with advancing a student whose true level is below π_o, and b is the loss for retaining a student whose true level exceeds π_o. The decision rule which minimizes expected loss in this situation is

to advance a student if his test score is such that

$$b\ \text{Prob}(\pi \geq \pi_0 | x, n) \geq a\ \text{Prob}(\pi < \pi_0 | x, n),$$

and to retain him otherwise. This comparison is equivalent to comparing the loss ratio a/b to the probability ratio $\text{Prob}(\pi \geq \pi_0 | x, n)/\text{Prob}(\pi < \pi_0 | x, n)$.

If a = b in our analysis, the decision procedure reduces to comparing the median of the posterior distribution with the specified criterion level. If the median is at least at this level, the student is advanced, otherwise he is retained. In this situation, the decision procedure is very similar to that used by Millman (1972). Though the procedure used by Millman is not Bayesian, it is equivalent to comparing with the mode (rather than the median) of the posterior distribution based on a uniform prior. Thus, in effect, the sampling theory approach gives equal weight to all equal intervals throughout the range of π; that is, effectively, to take π to be uniformly distributed apriori. This is seldom a reasonable prior specification. We might also remark that the formulation in Table 5 can be generalized to provide for differential utilities for correctly identifying true positives and true negatives as well as differential disutilities (or losses) for false positives and false negatives as is done in Table 5. To do this negative quantities (negative disutilities = utilities) would need to replace the zeros in Table 5, and a slightly more complicated analysis would be used.

It may be worthwhile to summarize the situation at this point. An instructor wishing to use test results in the context of Individually Prescribed Instruction should be ready to supply three kinds of information. First, a criterion level--the minimum degree of mastery required--must be set. In Individually Prescribed Instruction this seems to run from about

.70 to about .85. Second, prior knowledge of the student's true level of functioning must be translated into probability terms, namely a prior probability distribution for π. Typically, a carefully monitored program will be such as to suggest a prior probability distribution that assigns a probability of just more than .50 to the region above the criterion level. If this is not the case, the general efficacy of the program should be re-evaluated. A program that results in a much higher probability may be wastefully long and one that results in a lower probability may require strengthening. Finally, the relative losses associated with the two types of incorrect decisions must be assessed. A ratio of more than 1/1 is the rule (we are told) with ratios of 1.5/1 and 2/1 being common, and ratios as high as 3/1 not being rare.

It should be clear that all three of the above determinations will have an influence on the minimum necessary test length. As the criterion level approaches unity, the test must be longer in order to provide adequate information about a student's level of functioning in the neighborhood of the criterion. If prior probabilities of mastery are sufficiently high, very short tests become possible, but this is not and should not be the typical case. Finally, higher loss ratios require longer tests to allow the possibility of high posterior probability of mastery. We shall also see that greater test lengths are sometimes required because of the obvious restriction to integer valued sample sizes.

A Design For Test-Length Specification

The characteristics of the group of students being tested must now be considered as they relate to test-length specification. Each member

Table 6

Selected Prior Distributions For IPI Advancement Decisions

No.	Prior Distribution	Effective Prior Sample Size	Mean	Prob($\pi_\ell \leq \pi \leq \pi_u$)* .00–.70	.70–.75	.75–.80	.80–.85	.85–.90	.90–1.00
1	β(5.6, 2.4)	8	.70	.46	.12	.12	.12	.10	.08
2	β(6, 2)	8	.75	.33	.12	.13	.14	.13	.15
3	β(6.4, 1.6)	8	.80	.21	.10	.12	.15	.16	.26
4	β(6.8, 1.2)	8	.85	.12	.07	.09	.13	.17	.42
5	β(7.2, .8)	8	.90	.05	.04	.06	.09	.14	.62
6	β(7, 3)	10	.70	.46	.14	.14	.12	.09	.05
7	β(7.5, 2.5)	10	.75	.32	.13	.15	.15	.13	.12
8	β(8, 2)	10	.80	.20	.10	.14	.16	.17	.23
9	β(8.5, 1.5)	10	.85	.10	.07	.10	.14	.19	.40
10	β(9, 1)	10	.90	.04	.03	.06	.10	.16	.61
11	β(8.4, 3.6)	12	.70	.47	.15	.15	.12	.08	.03
12	β(9, 3)	12	.75	.32	.14	.16	.16	.13	.09
13	β(9.6, 2.4)	12	.80	.18	.11	.15	.18	.18	.20
14	β(10.2, 1.8)	12	.85	.09	.07	.11	.16	.20	.37
15	β(10.8, 1.2)	12	.90	.03	.03	.06	.11	.17	.60
16	β(10.5, 4.5)	15	.70	.47	.17	.16	.12	.06	.02
17	β(11.25, 3.75)	15	.75	.30	.16	.18	.17	.13	.06
18	β(12, 3)	15	.80	.16	.12	.17	.20	.19	.16
19	β(12.75, 2.25)	15	.85	.07	.07	.12	.18	.23	.33
20	β(13.5, 1.5)	15	.90	.02	.03	.06	.11	.19	.59

*Note: All entries have been rounded to two decimal places and smoothed so that the row totals add to 1.00.

of the group of students tested has been exposed to the same instruction program under identical local conditions. If a particular student is not considered atypical for this group, then our prior beliefs about his true level of functioning should closely reflect the true distribution of levels of functioning found in that group. Indeed, elaborate formal procedures for, effectively, bootstrapping a prior distribution using, for each examinee, the scores on the remaining m - 1 examinees are described by Novick, Lewis, and Jackson (1973). Thus, group characteristics, through their effect on our prior distributions, do affect test-length specification. If the average test score of the group is high (i.e., above the criterion level) and there is little variation among individuals, shorter tests become feasible.

Since, in practice, prior distributions will be based upon on-site experience, there will, of course, be different prior distributions for different sites. What we shall attempt to do here is to show what sample sizes will be required for a broad range of prior distributions and loss ratios. What we need to do now, therefore, is to consider certain combinations of prior distributions, criterion levels and loss ratios, and see what sample size will be adequate in each case.

For our analyses, we shall consider 20 different prior distributions for the level of functioning π, four specified criterion levels, and four loss ratios. For each criterion level, we shall consider all four loss ratios and four of the prior distributions. The four loss ratios we shall use are 1.5, 2.0, 2.5, and 3.0. The respective probabilities $P = \text{Prob}(\pi \geq \pi_0)$ required for advancement [given by setting $P/(1 - P)$ equal to the loss ratios, a/b] are .60, .67, .71, and .75. Thus, with a

loss ratio of 3.0, the posterior probability that the student's level of functioning is greater than the specified criterion level must be at least .75, if he is to be advanced.

The twenty prior probability distributions we shall be considering are given in Table 6 where they have been grouped in blocks of five, with each block having a distribution with the respective mean values .70, .75, .80, .85, and .90. The blocks differ with respect to the concentration of the prior distributions. Within block, the distributions differ with respect to their mean values. Note that in the first block the arguments of each Beta distribution sum to 8, e.g., 5.6 + 2.4 = 8. This indicates that the amount of prior information contained in each of these distributions is equivalent to what would be gained from a test containing eight items. If given one of these prior distributions and some criterion level and loss ratio, we specify an eight-item test, our posterior distribution will contain information equivalent to that contained in 16 observations. This contrasts with the classical procedure which uses no prior information. It is this increment in information that is equivalent to prior observations which permits a reduction in test length when a Bayesian procedure is used.

The first problem in doing an analysis is that of selecting a reasonable prior distribution. For the present application, we would first need to ask ourselves what we would expect to find as the mean level of functioning in our posttest group. With a specified criterion level of .70, we might hope for a mean level of functioning of .70. Thus, we would have people in training until such time as we would "expect" them to be qualified. Since loss ratios are typically greater than one, some overtraining may be thought to be useful, but as we shall see, excessive overtraining may be wasteful.

Suppose, for concreteness, that we believe the mean population level of functioning to be .70. Distributions 1, 6, 11, and 16 satisfy this condition, and, hence, we may choose from among these. We note that these distributions are in an increasing order of tightness, as may most conveniently be seen in the probability assignment given in the last column, to the interval (.90, 1.00). These probabilities are respectively .08, .05, .03, and .02. We need to ask ourselves which of these values seems most reasonable, and this then will give us some preference among these prior distributions. We might consider the relative weight of prior information assumed by each prior distribution (8, 10, 12, and 15 equivalent prior observations, respectively), and this should help to narrow our focus to one or two adjacent prior distributions for this, or any other application. Since the authors of this paper cannot know what an appropriate prior distribution will be in applications they have not seen, it will be most helpful, we think, to work out sample size allocations for several prior distributions and leave the final selection to be made "in the field". We believe that the prior distributions, loss ratios, and specified criterion levels used here are typical of those found in practice, and, therefore, that the specific results we shall obtain will be useful. However, if other combinations present themselves, we believe that the general methodology that we are demonstrating should be adequate to the problem. Actually we shall find that most of our specifications are very robust with respect to the choice of prior distribution within the range we have considered.

Some Specific Test Length Recommendations

In Table 7, we give recommended sample sizes and minimum advancement scores for $\pi_o = .70$, $(a/b) = 1.5, 2.0, 2.5, 3.0$ and prior distributions 1, 6, 11, and 16. The values that we have settled on for the body of

Table 7

Recommended Sample Sizes and Advancement Scores

$\pi_o = .70$

Prior Distribution	$\xi(\pi)$	Loss Ratio 1.5 (.60)	2.0 (.67)	2.5 (.71)	3.0 (.75)
$\beta(5.6, 2.4)$[1]	(.70)	6/8(.62)	10/13(.70)	11/14(.74)	12/15(.78)
$\beta(7, 3)$	(.70)	6/8(.61)	10/13(.69)	11/14(.73)	12/15(.77)
$\beta(8.4, 3.6)$	(.70)	6/8(.61)	10/13(.68)	11/14(.72)	12/15(.76)
$\beta(10.5, 4.5)$	(.70)	9/12(.62)[2]	10/13(.67)	11/14(.71)	12/15(.75)
		General Recommendations			
		6/8(75%)	10/13(77%)	11/14(79%)	12/15(80%)

[1] Apriori, Prob($\pi \geq .70$) for each of the four prior distributions is .54, .54, .53, and .53.

[2] For 6/8, Prob($\pi \geq .70$) = .598.

this table are not, in every instance, optimum in any statistical sense, though we are confident that the risks associated with these decision rules are in every case insignificantly different from the risks of the optimum procedures. In selecting values for this table we have sought sample sizes and minimum advancement scores that would be very efficient over a wide range of prior distributions. That we have been successful in this endeavor is confirmed by our ability to give general recommendations that hold throughout the range of prior distributions studied. Actually in only one instance have we cheated (see Footnote 2, Table 7), but again the increase in expected loss will be trivial. We would also note that the required percentage correct and the number of required observations increases as the loss ratio increases, which "makes sense" on intuitive grounds.

A rough indication of the near optimality of any of the individual specifications can be gained from the closeness of the aposteriori probability (indicated in parentheses following the specification) with the value required by the particular loss ratio (given in parentheses at the top of the column). Thus, with the prior distribution $\beta(7, 3)$, the decision rule "six out of eight", abbreviated 6/8, leads to the aposteriori distribution $\beta(13, 5)$ and to $\text{Prob}(\pi > .70) = .61$ which is just .01 greater than the required level .60 for the loss ratio 1.5 (1.5 to 1). In this instance, the specified decision rule may be very good. On the other hand, consider the prior distribution $\beta(5.6, 2.4)$. Here the rule 11/14 leads to a value .74 when only .71 is required for a 2.5 to 1 loss ratio.

Actually, the specification 8/10 is somewhat better giving a posterior probability of .729. Also for the prior distribution $\beta(7, 3)$, the posterior probability with 8/10 is .718. With the loss ratio 2.0/1 and with the prior $\beta(5.6, 2.4)$, the rule 7/9 leads to the posterior probability .68 as compared to desired value of .67. In every case where we have specified an "almost best" decision rule, the result has been an increase in the specified sample size and the purpose has been to obtain uniformity of specification over a reasonably wide range of amounts of prior information. Considering our general ignorance concerning what might be an appropriate prior distribution in specific applications, the specifications we have given should be the more generally useful.

Another indication of how good a particular specification is can be inferred from the closeness of the percentage correct required by the advancement rule to the specified criterion level. Clearly, if the percentage required by the advancement rule is very much larger than the specified criterion level, a large percentage of qualified students will be retained and this is undesirable, particularly for small loss ratios. For large loss ratios, this is less important and hence higher advancement ratios can, and will need to be tolerated. This feature is exhibited in Table 7, where the advancement ratios increase with increasing loss ratios. One can, of course, keep the advancement ratio down very close to the specified criterion level even for higher loss ratios, but only by having much larger sample sizes. For example with the prior distribution $\beta(5.6, 2.4)$ the specified criterion level $\pi_0 = .70$ and the loss ratio 2.0, the advancement ratio 72/100 is satisfactory since $\text{Prob}(\pi > .70 \mid 72/100) = .675$, but the indicated sample size is unacceptable.

Table 8

Recommended Sample Sizes and Advancement Scores

$\pi_o = .75$

Prior Distribution	$\xi(\pi)$	Loss Ratio 1.5 (.60)	2.0 (.67)	2.5 (.71)	3.0 (.75)
$\beta(6, 2)$[1]	(.75)	8/10(.65)	16/20(.70)	17/21(.74)	18/22(.77)
$\beta(7.5, 2.5)$	(.75)	8/10(.64)	16/20(.69)	17/21(.73)	18/22(.76)
$\beta(9, 3)$	(.75)	8/10(.63)	16/20(.69)	17/21(.72)	18/22(.75)
$\beta(11.25, 3.75)$	(.75)	8/10(.62)	16/20(.68)	17/21(.71)	19/23(.77)[2]
			General Recommendations		
		8/10(80%)	16/20(80%)	17/21(81%)	18/22(82%)

[1]Apriori, Prob($\pi \geq .75$) = .56, .55, .55, and .54, respectively, for the four prior distributions used in Table 8.

[2]For 18/22, Prob($\pi \geq .75$) = .744.

Table 9

Recommended Sample Sizes and Advancement Scores

$\pi_o = .80$

Prior Distribution	$\xi(\pi)$	Loss Ratio 1.5 (.60)	2.0 (.67)	2.5 (.71)	3.0 (.75)
$\beta(6.4, 1.6)$[1]	(.80)	6/7(.66)	7/8(.70)	17/20(.72)	19/22(.78)
$\beta(8, 2)$	(.80)	6/7(.65)	7/8(.69)	17/20(.72)	19/22(.77)
$\beta(9.6, 2.4)$	(.80)	6/7(.64)	7/8(.68)	17/20(.71)	19/22(.76)
$\beta(12, 3)$	(.80)	6/7(.63)	7/8(.67)	18/21(.73)[2]	19/22(.75)
			General Recommendations		
		6/7(86%)	7/8(83%)	17/20(85%)	19/22(86%)

[1]Apriori, Prob($\pi \geq .80$) = .57; for 8/10, Prob($\pi \geq .80$) = .55; for 16/20, Prob($\pi \geq .80$) = .54; for 8.5/10, Prob($\pi \geq .80$) = .67; for 8.3/10, Prob($\pi \geq .80$) = .62; for 9/10, Prob($\pi \geq .80$) = .78.

[2]For 17/20, Prob($\pi \geq .80$) = .70.

Note that for each of the prior probabilities used in Table 7, Prob($\pi \geq .70$) > .50. Thus, on an apriori basis, advancement would be indicated with a loss ratio 1.0. This will generally be true for the prior distributions we shall be adopting for our analyses. The point is that loss ratios of 1.0 are not (we are told) typical of IPI applications, and if test lengths are to be kept reasonable it will be necessary to use training programs that give mean output at or above the criterion level.

There has been a definite tendency in IPI to require relatively high advancement ratios; typically, the value .85 is used. One might well speculate whether this is a function of a high loss ratio combined with a desire for a short test length, or whether it really reflects a perceived need for a high criterion level. (For example an advancement ratio of 6/7 with the prior distribution $\beta(5.6, 2.4)$ would yield with $x = 6$ a posterior Prob($\pi > .70$) = .77 which would be just right with a loss ratio of 3.0.) The authors of this paper do not know the answer to this question, but hope that those within IPI will want to consider it carefully. Only through such serious consideration can the test length problem be "solved".

Some recommended test lengths for $\pi_o = .75$ and four prior distributions with $\mathcal{E}(\pi) = .75$ are given in Table 8. Again we have been able to specify one generally satisfactory advancement ratio for each of the four loss ratios. We note that the required test lengths for $\pi_o = .75$ are rather larger than for $\pi_o = .70$. In Table 8, we find very short required test lengths for a 1.5 loss ratio and rather long ones for loss ratios of 2.0, 2.5, and 3.0.

In Table 9, we provide recommendations for $\pi_o = .80$ when $\mathcal{E}(\pi) = .80$. The results here parallel those of Table 8, except that the advancement ratios are very high as compared to the criterion levels. This is

relatively unsatisfactory. In Footnote 1 to Table 9, we indicate the formal results for the prior distribution $\beta(6.4, 1.6)$ and the sample result "8.5" correct and "1.5" incorrect and also for "8.3" correct and "1.7" incorrect. These provide very nice results for loss ratios of 2.0 and 1.5, respectively. Unfortunately, these are unobtainable sample results. This demonstrates that in part, large required test lengths may sometimes be due to the discreteness, and hence, discontinuity of our possible experimental outcomes. This also suggests that the precise specification of the advancement rules may be highly sensitive to the mean value of the prior distribution even if it is proving to be relatively insensitive to the total amount of information contained in the prior distribution, which is indicated by the sum of the two parameters of the Beta distribution.

For example, given the prior distribution $\beta(6.4, 1.6)$ and the impossible sample result $x = 8.3$, $n = 10$, we have the posterior distribution $\beta(14.7, 3.3)$ which, as we indicated previously, gives $\text{Prob}(\pi > .80) = .62$ which suggests that the advancement ratio 8.3/10 might be very favorable with a loss ratio of 1.5. But suppose we had just a slightly different prior distribution, namely, $\beta(6.7, 1.3)$ with $\mathcal{E}(\pi) = .84$, then the sample result $x = 8$, $n = 10$ would yield the posterior distribution $\beta(14.7, 3.3)$ and thus, for the reasons given above, indicate that the advancement ratio 8/10 might be attractive. This advancement ratio is clearly more attractive than the ratio 6/7, despite the fact that it requires three additional items, because this ratio $8/10 = 80\%$ is closer to the criterion level than is the advancement ratio $6/7 = 86\%$.

Because of this relatively high dependence of the results on the expected value of the prior distribution, it seems important to attempt some study of the variation of our results as a function of changes in

Table 10

Recommended Sample Sizes and Advancement Scores

$\pi_o = .80$

Prior Distribution	Loss Ratio 1.0 (π)	1.5 (.60)	2.0 (.67)	2.5 (.71)	3.0 (.75)
$\beta(6.8, 1.2)$[5]	(.85)	8/10(.64)	9/11(.69)	10/12(.72)[1]	11/13(.76)
$\beta(8.5, 1.5)$	(.85)	8/10(.66)	9/11(.70)	10/12(.73)[2]	11/13(.76)
$\beta(10.2, 1.8)$	(.85)	8/10(.67)	9/11(.71)	9/11(.71)[3]	11/13(.77)
$\beta(12.75, 2.25)$	(.85)	8/10(.69)	9/11(.72)	9/11(.72)[4]	11/13(.78)
		General Recommendations			
		8/10(80%)	9/11(82%)	10/12(83%)	11/13(85%)

[1] For 5/6, Prob($\pi \geq .80$) = .72.

[2] For 5/6, Prob($\pi \geq .80$) = .73.

[3] For 10/12, Prob($\pi \geq .80$) = .74.

[4] For 10/12, Prob($\pi \geq .80$) = .75.

[5] For the four prior distributions, the apriori probabilities of $\pi \geq .80$ are .72, .73, .74, and .75. With these prior distributions and with 7/10, the posterior probabilities of $\pi \geq .80$ are .41, .43, .46, and .48.

our prior distribution. For this reason, we have in Table 10 redone our sample size recommendations under the assumption that the mean of our prior distribution is .85 instead of .80.

Surely the practitioner will find the sample size recommendations of Table 10 to be attractive. Apparently with these prior disbributions, test lengths need be no greater than 13 for any of the listed loss-ratios. With the prior distributions having $\xi_o^o(\pi) = .80$, a sample size of 22 is required when the loss ratio is 3.0.

What is happening is that we are beginning with fairly strong beliefs that $\pi \geq \pi_o$ so that not much data, in confirmation, is required even for high loss ratios. In fact, even on an apriori basis, an advancement decision would be made for all loss ratios up to and including 2.5. Indeed, we see that the function of the sample data here is to provide the possibility of obtaining some information that might change the decision to retention. For example, an observed performance ratio of 10/13 with the prior distribution $\beta(6.8, 1.2)$ would give aposteriori $\text{Prob}(\pi \geq .80) = .72$, and hence, the student would be retained if the loss ratio were 3.0 (see also Footnote 5, Table 10).

We believe that the comparison of the specifications in Tables 9 and 10 have important implications for IPI management. When loss ratios are high, it may well be highly advantageous to strengthen the training program to the extent that the mean output is well above the specified criterion level. This will make it possible to use short tests or, alternatively will generally reduce the risk of incorrect classification. This will, of course, be more expensive, and this investment must be balanced out against the reduction in the cost of testing and the reduction in the expected loss due to incorrect decision. The final Table, Table 11, looks

Table 11

Recommended Sample Sizes and Advancement Scores

$\pi_o = .85$

Prior Distributions	$\xi^0(\pi)$	Loss Ratio 1.5 (.60)	2.0 (.67)	2.5 (.70)	3.0 (.75)
$\beta(6.8, 1.2)$[1]	(.85)	7/8(.62)	9/10(.70)	17/19(.73)	18/20(.76)[3]
$\beta(8.5, 1.5)$	(.85)	7/8(.62)	9/10(.69)	17/19(.72)	19/21(.77)
$\beta(10.2, 1.8)$	(.85)	7/8(.61)	9/10(.68)	17/19(.72)	19/21(.76)
$\beta(12.75, 2.25)$	(.85)	7/8(.60)	9/10(.67)	17/19(.71)[2]	19/21(.75)
		General Recommendations			
		7/8(87.5%)	9/10(90%)	17/19(89%)	19/21(90%)

[1]The apriori probabilities for $\pi \geq .85$ are .59, .58, .58, and .57.

[2]For 10/11, Prob($\pi > .85$ = .695).

[3]For 19/21, Prob($\pi > .85$ = .78).

very much like Table 9 as far as test lengths are concerned. Here again some robust length assignments are obtained, though again, the lengths for the high loss ratios border on being discomforting. This can be corrected by training to an average level of functioning of .90. With the prior distribution $\beta(7.2, 8)$, we find that $\text{Prob}(\pi \geq .85) = .76$, apriori. Observing 6/7 yields $\text{Prob}(\pi \geq .85) = .70$, while 5/7 yields a value of .41. Observing 8/9 yields .77, while 7/9 yields .493. Thus, clearly, very short test lengths are again possible if the students are trained to a sufficiently high average standard.

Some Summary Remarks

The test length recommendations given in this paper are meant to be taken seriously and hopefully they will soon be adopted on a provisional and experimental basis, so that more experience can be gained while some of the theoretical and substantive issues raised in this paper are debated. The questions of level of functioning required to define mastery and the relative losses incurred in making false positives and false negative decisions require serious discussion and concensus. We also need to get some clear picture of what kinds of distributions of outcomes are to be expected as this determines the amount of prior information available in making individual assessments. This third issue is, as we have indicated, intimately related to the expected level of functioning that is sought in the group being trained. Hopeful and possible outcomes of such discussions could be a consensus that:

1. In most situations a level of functioning of something less than .85 is satisfactory. A value as low as .75 would be highly desirable. This could be accomplished by redefining the task domain slightly to eliminate very easy items.

2. Training should be carefully monitored so that expected group performance will be just slightly higher than the specified criterion level. This will keep training time and testing time relatively low.
3. The program should be structured so that very high loss ratios are not appropriate. That is to say, individual modules should not be overly dependent on preceding ones.

One problem that does not arise with Bayesian methods is any complication if sequential methods are used. Items can simply be administered until it is clear that a student will definitely, or cannot possibly, attain the minimum advancement score. Thus with a minimum advancement score of 8/10, testing can cease as soon as light successes or three failures are observed.

Two issues have been treated in a rather gross way in this paper and on these important issues further research needs to be done. First it must be recognized that while the threshold loss function we have adopted here is a better approximation to reality than, for example, Livingston's criterion centered squared-error loss (see Hambleton and Novick, 1973), it is only a gross approximation to be used while better and more complicated approximations are being investigated. Three that immediately come to mind are:

1. A threshold loss function with an indifference region in which there is zero loss for false positive or false negative errors.
2. A negative squared-exponential loss used with the root arcsine transformation parameter

$$\gamma = \sin^{-1}\sqrt{\pi} \quad .$$

3. A cumulative Beta distribution loss function.

We expect that these loss functions will give somewhat different and surely better length specifications than those obtained here, but the overall decrease in expected loss may or may not be great. We should also remark that these recommendations are specifically made for first time through decisions. We have yet to consider the problem of decisions for students repeating a unit.

Finally, we would remark that one of the important issues that we identified at the outset of this paper has been handled in a most casual and informal manner. To do other than this would have enormously complicated the analysis and delayed substantially the appearance of our recommendations. We refer explicitly to the premium on testing time within the instructional process and implicitly to an implied trade-off between training and testing time. A completely general analysis would consider an available time T and an allocation of T into instruction and testing times $i + t = T$, so as to maximize a payoff function which would have a (possibly differential) positive payoff for each module successfully completed, and a (differential) negative payoff for an incorrect decision of either type. We are reluctant to undertake such a sophisticated analysis until such time as the operating conditions of IPI are more clearly defined.

For the present paper we have implicitly adopted some guidelines which effectively say that it is very desirable to have test lengths of 12 or less, tolerable but undesirable to have test lengths as high as 20 and discomforting to have tests that are longer than this. We have also taken the position that a decision should not be made on the basis of prior and

collateral information alone but that mastery must be confirmed by a test that permits demonstration of nonmastery. As in all of the judgmental decisions made in this paper we have been guided by counsel from experienced IPI personnel, particularly Richard Ferguson and Anthony Nitko to whom we are much indebted. The value of this paper will largely be determined by the quality of the discussion engendered by it among such people.

REFERENCES

Hambleton, R., & Novick, M. R. Toward an integration of theory and method for criterion-referenced tests. Journal of Educational Measurement, 1973.

Lewis, C., Wang, M., & Novick, M. R. Marginal distributions for the estimation of proportions in m groups. ACT Technical Bulletin No. 13. Iowa City, Iowa: The American College Testing Program, 1973.

Millman, J. "Determining Test Length, Passing Scores and Test Lengths for Objective-Based Tests." Los Angeles: Instructional Objectives Exchange, 1972.

Novick, M. R. High school attainment: An example of a computer-assisted Bayesian approach to data analysis. International Statistical Review, 1973, 41, 264-271.

Novick, M. R., & Jackson, P. H. Statistical Methods for Educational and Psychological Research. New York: McGraw-Hill, 1974.

Novick, M. R., Lewis, C., & Jackson, P. H. The estimation of proportions in m groups. Psychometrika, 1973, 38, 19-46.

Wang, M. Tables of constants for the posterior marginal estimates of proportions in m groups. ACT Technical Bulletin No. 14. Iowa City, Iowa: The American College Testing Program, 1973.